Motivation:

Time Allocation:

A significant amount of my time is spent on YouTube. What hours do I spend most of my time on this platform? I wonder if YouTube has a negative effect on my sleep pattern?

Content Interaction:

Investigating which categories of YouTube content I engage with most frequently. Is there a category that has pressure over the others, or are the topics evenly distributed? Do my video preferences follow a pattern, are they predictable, if how much of accuracy?

Duration of Engagement:

Quantifying the time spent per content category on YouTube. Are the videos I watch instructive?

Anticipated Outcomes:

Learning from the analysis to positively influence my YouTube habits and improve digital well-being.

Data Source

Google Takeout Service:

Utilized to extract my complete YouTube activity history.

YouTube API:

Assisted in retrieving detailed metadata, such as video categories and durations.

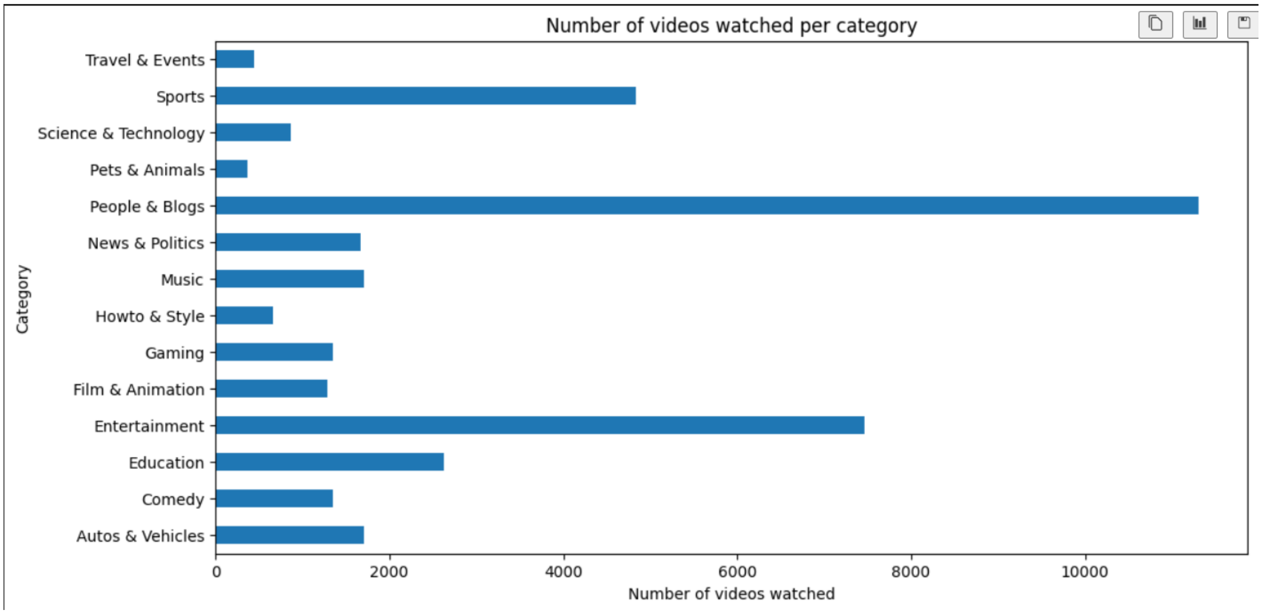
Data Composition:

The dataset comprises my YouTube click history, including timestamps and video titles.

Data Cleaning:

Performed to remove irrelevant information, ensuring a focus on meaningful data for analysis. Removed the samples with incomplete information for sake of consistency and the ml models that I used require valid data.

Data Analysis:



Most Viewed Categories:

The data highlights 'Entertainment' and 'People & Blogs' as the most viewed categories, with the highest number of videos watched.

Educational Content:

The 'Education' and 'Science & Technology' categories show a moderate level of engagement, indicating that I use it for entertainment purposes rather than informative content.

After observing a pattern of consistency in the categories of videos I was frequently watching, I was inspired to leverage the power of machine learning to anticipate the topic of my future video engagement. To achieve this, I conceptualized a predictive model that would consider a series of variables known to influence viewing choices.

In this project, I employed various techniques for data analysis and predictive modeling:

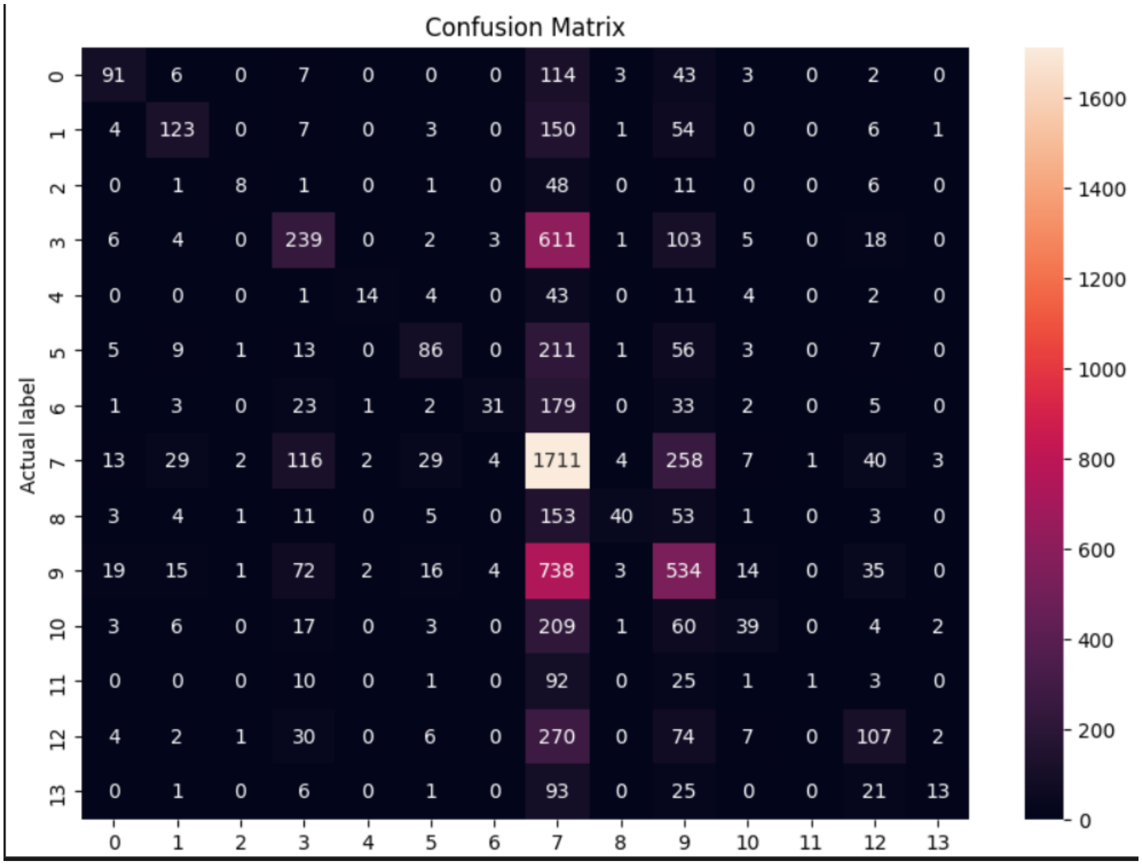
Feature Engineering: Created time-based features and identified the most-watched categories by week and day. I also kept track of the duration, category and timestamps of the previous video because I believe there is a sequential pattern.

Data Preparation: Shifted data to predict the next video's category, ensuring the model could not inadvertently 'peek' into the future (preventing look-ahead bias).

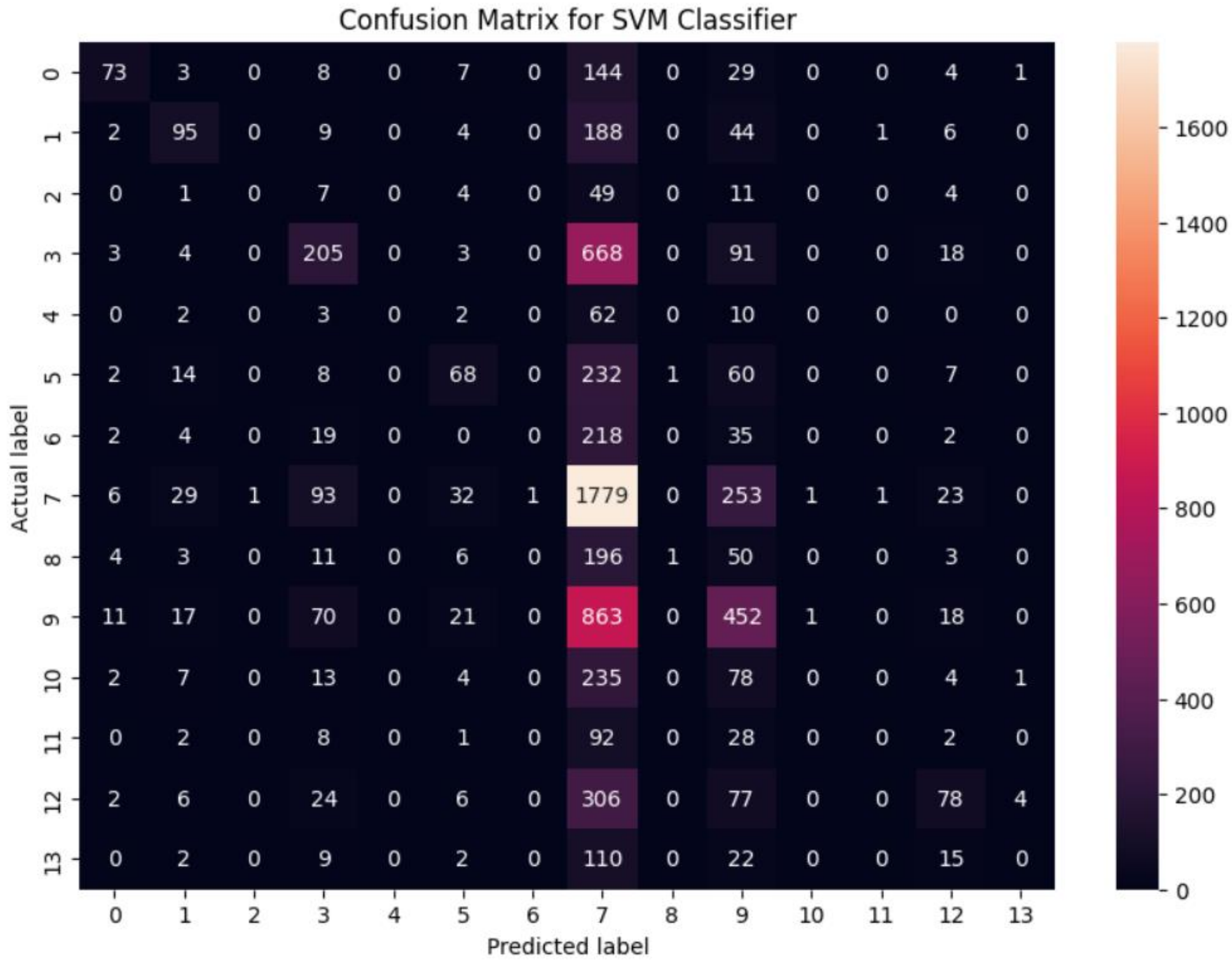
Modeling:

Random Forest: Deployed for its effectiveness in classification tasks, with hyperparameters optimized via RandomizedSearchCV. This approach was chosen due to the suspicion of a wide range of close-to-optimal models generated by different hyperparameters, and because Random Forest has many hyperparameters to consider.

SVM: Selected for its superior performance in high-dimensional spaces and refined using GridSearchCV within a scaling pipeline. This was because there are not really many hyperparameter combinations, so optimizing via grid search was deemed beneficial.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	91	6	0	7	0	0	0	114	3	43	3	0	2	0
1	4	123	0	7	0	3	0	150	1	54	0	0	6	1
2	0	1	8	1	0	1	0	48	0	11	0	0	6	0
3	6	4	0	239	0	2	3	611	1	103	5	0	18	0
4	0	0	0	1	14	4	0	43	0	11	4	0	2	0
5	5	9	1	13	0	86	0	211	1	56	3	0	7	0
6	1	3	0	23	1	2	31	179	0	33	2	0	5	0
7	13	29	2	116	2	29	4	1711	4	258	7	1	40	3
8	3	4	1	11	0	5	0	153	40	53	1	0	3	0
9	19	15	1	72	2	16	4	738	3	534	14	0	35	0
10	3	6	0	17	0	3	0	209	1	60	39	0	4	2
11	0	0	0	10	0	1	0	92	0	25	1	1	3	0
12	4	2	1	30	0	6	0	270	0	74	7	0	107	2
13	0	1	0	6	0	1	0	93	0	25	0	0	21	13



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	73	3	0	8	0	7	0	144	0	29	0	0	4	1
1	2	95	0	9	0	4	0	188	0	44	0	1	6	0
2	0	1	0	7	0	4	0	49	0	11	0	0	4	0
3	3	4	0	205	0	3	0	668	0	91	0	0	18	0
4	0	2	0	3	0	2	0	62	0	10	0	0	0	0
5	2	14	0	8	0	68	0	232	1	60	0	0	7	0
6	2	4	0	19	0	0	0	218	0	35	0	0	2	0
7	6	29	1	93	0	32	1	1779	0	253	1	1	23	0
8	4	3	0	11	0	6	0	196	1	50	0	0	3	0
9	11	17	0	70	0	21	0	863	0	452	1	0	18	0
10	2	7	0	13	0	4	0	235	0	78	0	0	4	1
11	0	2	0	8	0	1	0	92	0	28	0	0	2	0
12	2	6	0	24	0	6	0	306	0	77	0	0	78	4
13	0	2	0	9	0	2	0	110	0	22	0	0	15	0

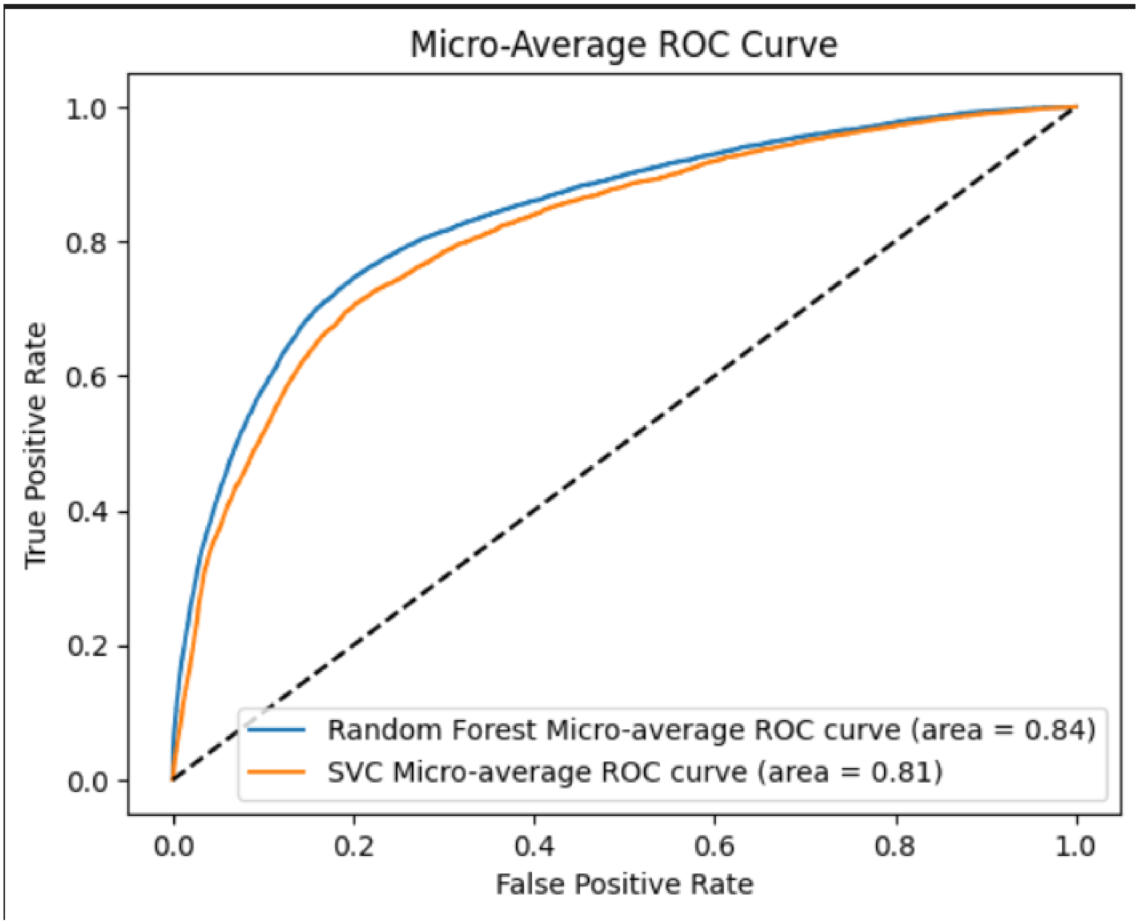
Diagonal Cells: Indicate correct predictions; lighter shading represents higher accuracy per class.

Off-Diagonal Cells: Represent incorrect predictions; for example, class 7 shows confusion with other classes.

Desired Outcome: Darker diagonal with lighter off-diagonal cells, signifying strong model accuracy.

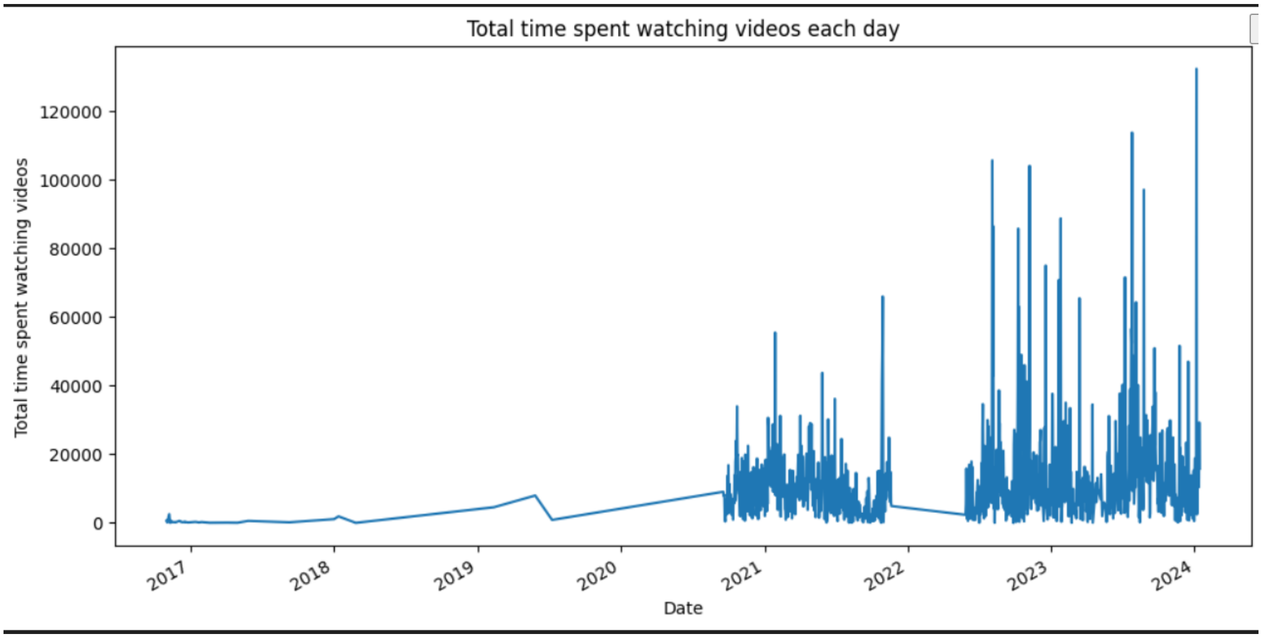
Observations: High accuracy for some classes (e.g., third row), whereas others like class 7 show more misclassifications.

Due to the abundance of data from certain categories, predicting labels outside of these categories becomes challenging. However, Random Forest appears to handle this issue more effectively. I believe the reason is its tree-based structure, which facilitates easier discovery of interactions between variables.



The Random Forest model's ROC curve positioned closer to the top-left corner of the plot indicates its superior performance in distinguishing between classes compared to the SVC model. This positioning suggests a higher true positive rate with a lower false positive rate, signifying that the Random Forest model effectively maximizes correct classifications while minimizing incorrect ones for this dataset.

Findings:

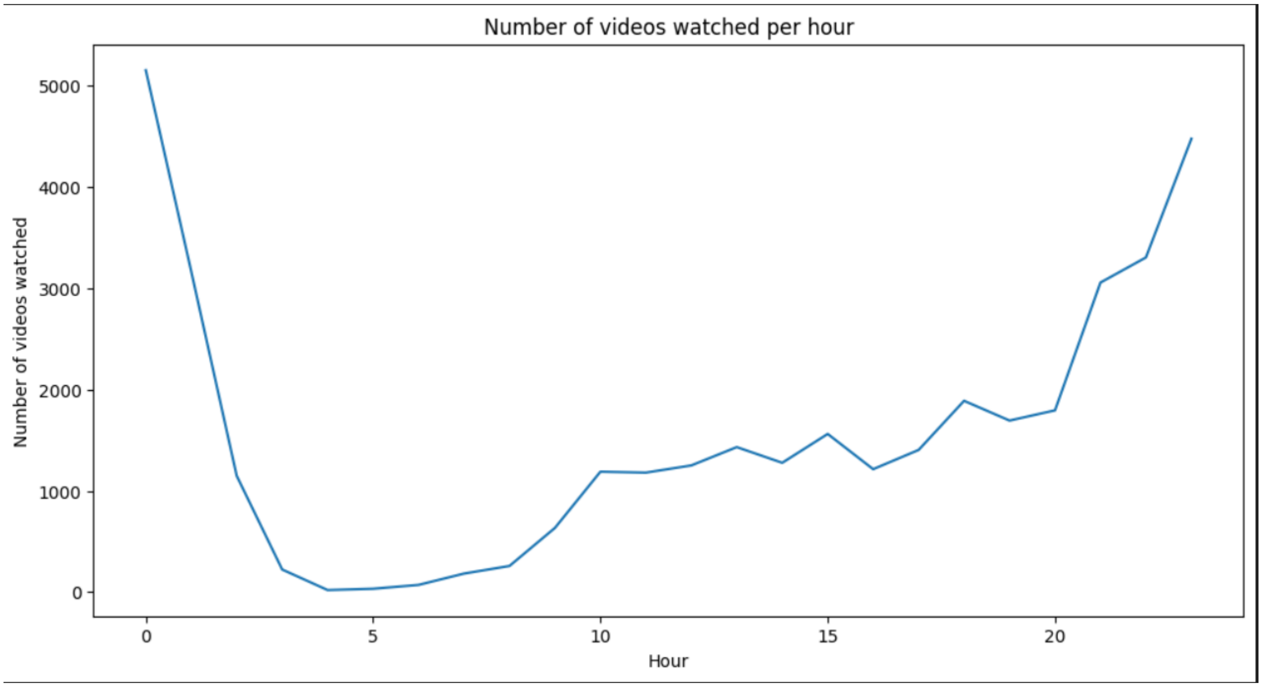


Pre-Pandemic Usage:

Initially, my YouTube usage was relatively stable and moderate, as indicated by the lower and consistent data points before 2020.

Pandemic Effect:

There is a discernible surge in YouTube consumption beginning around the onset of the global pandemic in early 2020.



Nighttime Activity:

There is a pronounced increase in the number of videos watched during the late evening hours, suggesting that my YouTube usage intensifies at night.

Potential Impact on Sleep:

This trend of increased nighttime activity could be affecting my sleep schedule, potentially leading to later bedtimes.

Early Hours Drop:

After midnight, there is a sharp decline in activity, which coincides with the time I likely go to sleep.

Daytime Viewing:

During the day, my YouTube usage remains relatively steady with slight increases in the afternoon hours.

Limitations and Future Work

Limitations:

API Data Fields: The current dataset has limitations due to the lack of certain fields in the API. For example, the inclusion of 'video watch percentage' could significantly enhance the model by providing insights into not just which videos were clicked, but also how engaging each video was, which is a more robust indicator of viewer interest. A 'better categorization' field would allow for more precise predictions and a deeper understanding of content preferences.

Additional Useful Field: An 'audience interaction data' field, including likes, comments, and shares, would be beneficial. This data could help understand the level of engagement and sentiment towards the videos, thereby refining the model's ability to predict future viewing behavior based on interaction levels, which often correlate with content relevance and viewer satisfaction.

Future Work:

Despite the limitations, there is no immediate plan for future work on this project. The current scope of the project has provided sufficient insights, and the machine learning model has achieved a satisfactory level of accuracy in predicting video categories. Therefore, the project in its current state is considered complete. However, should there be significant updates or changes in the YouTube API offering more detailed data fields, there may be an opportunity to revisit and enhance the model for even more nuanced analysis and predictions.